PREDIKSI *DELAY* PENERBANGAN MENGGUNAKAN GAUSSIAN PROCESS REGRESSION

[Tugas Besar MA4282 Kapita Selekta Statistika II Fakultas Matematika dan Ilmu Pengetahuan Alam Institut Teknologi Bandung]

LATAR BELAKANG

100,000++ penerbangan setiap harinya di seluruh dunia (ICAO, 2019)

Delay penerbangan bukan kejadian langka

Bagaimana memprediksi kemungkinan terjadinya *delay*?

Metode: *Dynamic Gaussian Process Regression*

Tujuan:
Passengers dapat menggunakan model untuk memilih waktu terbaik untuk melakukan penerbangan.

GAUSSIAN PROCESS REGRESSION

- **Gaussian Processes (GP)**: Generalisasi distribusi Gaussian multivariat. Asumsikan *f(x)* adalah proses pada input *x*, maka
  f(x) ~ GP(0, k(x,x')).
- **Gaussian Processes Regression (GPR):** GP dengan pendekatan regresi nonparametrik Bayesian.

- **Distribusi Posterior**:
  Misalkan data observasi $\{x_{o,i}, y_{o,i}\}_{i=1}^{n}$ maka *f* pada titik baru x_* mengikuti distribusi Gaussian dengan rataan dan variansi

$$\mu_* = k(x_*, x_o)(K_{o,o} + \sigma_o^2 I)^{-1} y_o$$

$$\sigma_*^2 = k(x_*, x_*) - k(x_*, x_o)(K_{o,o} + \sigma_o^2 I)^{-1} k(x_o, x_*)$$

dengan $K_{o,o}|_{i,j} = \mathbf{K}|_{i,j} = k(x_{o,i}, x_{o,j})$

- **Optimisasi *Hyperparameter:***
  Memaksimumkan log-likelihood

$$\pounds(\theta) = -\frac{1}{2} y^T K^{-1} y - \frac{1}{2} log(|K|) - \frac{n}{2} log(2\pi)$$

DYNAMIC GPR

Fungsi kernel ARD:

$$k(x, x') = \varphi_1^2 exp(-(x - x')^T P^{-1} (x - x'))$$

$$P = diag(\varphi_2^2, ..., \varphi_{n_\varphi}^2)$$

Dengan menggunakan K-means diperoleh basis

$$\{c_{b,i}\}_{i=1}^{M}$$

sehingga $\{(c_i, u_i)\}_{i=1}^{M}$ adalah bagian dari *GP(0,k)* dengan

$$\mathbf{u} = \begin{pmatrix} u_1 & u_2 & \cdots & u_M \end{pmatrix}^T \sim N(\mathbf{m}, \mathbf{S})$$

sehingga diperoleh

$$u_n(x_*) \mid m_n, s_n \sim \mathcal{N}\big(\mu_n(x_*), \Sigma_n(x_*, x_*)\big)$$

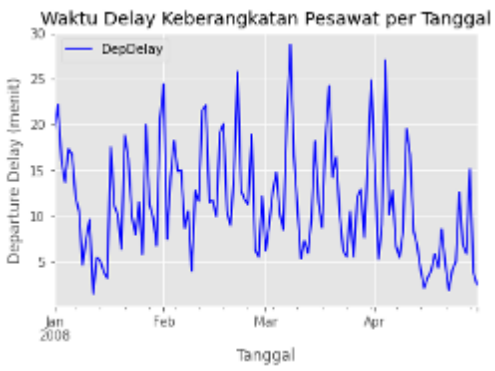
dengan

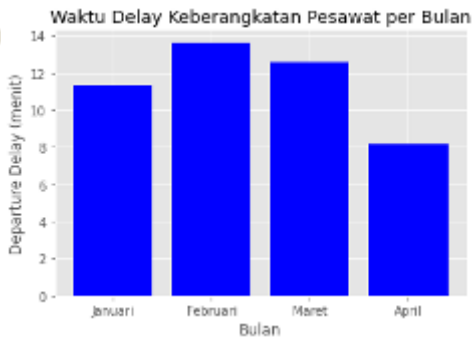
$$\mu_n(x) = k(x, c_b) K_{c_b,c_b}^{-1} m_n$$

dan

$$\Sigma_n(x, x') = k(x, x'; \theta_n) - k(x, c_b) K_{c_b,c_b}^{-1} k(c_b, x') + k(x, c_b) K_{c_b,c_b}^{-1} S_n K_{c_b,c_b}^{-1} k(c_b, x')$$

1 Pada bulan Maret 2008, rerata waktu keberangkatan mengalami keterlambatan selama hampir 30 menit.





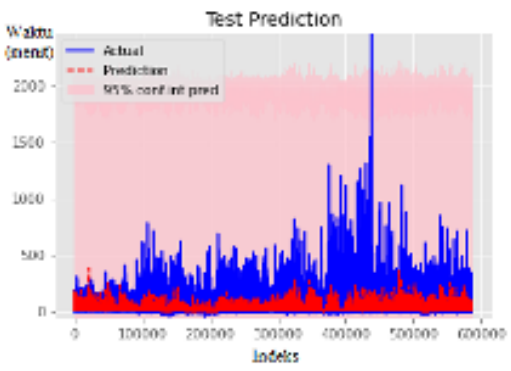
2 Waktu *delay* keberangkatan pesawat tertinggi tercapai pada bulan Februari dan hari Jumat.

ANALISIS DATA

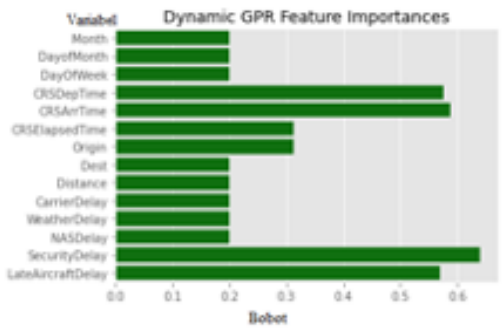
Data yang digunakan adalah data penerbangan maskapai (*airlines*) US pada tahun 2008, yang memiliki 29 *features*. Dengan melakukan analisis data diperoleh informasi seperti yang tertera di sebelah kiri.

EKSPERIMEN

Eksperimen dilakukan dengan menggunakan observasi pada bulan Jan-Mar sebagai data *train* dan observasi pada bulan Apr sebagai data *test*.



Dengan 25 titik basis, diperoleh hasil prediksi model pada data test dengan
Akurasi: 84.35%
MSE: 941.8354



Variabel yang memegang peran penting dalam memprediksi variabel DepDelay antara lain:

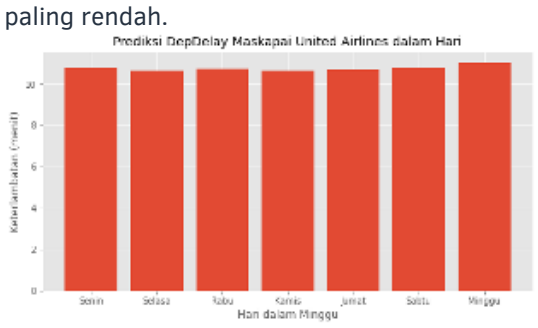
- SecurityDelay,
- CRSArrTime,
- CRSDepTime, dan
- LateAircraftDelay

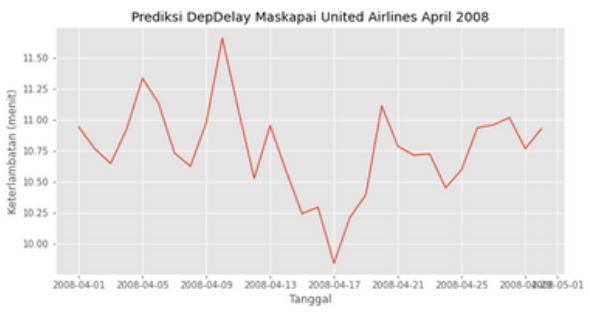
Waktu keterlambatan keberangkatan paling rendah secara rata-rata akan terjadi pada 13-17 April.

<u>Hasil forecast model pada bulan April 2008</u>
Hari Rabu menjadi hari dengan waktu keterlambatan keberangkatan paling rendah

<u>Prediksi keterlambatan keberangkatan maskapai United Airlines</u>

Hari Selasa, Kamis, dan Jumat menjadi hari dengan waktu keterlambatan keberangkatan paling rendah.





Waktu keterlambatan keberangkatan paling rendah secara rata-rata akan terjadi pada 17 April.

KESIMPULAN

Berdasarkan hasil penelitian yang dilakukan didapatkan hasil

- Pada bulan April 2008, untuk seluruh maskapai, dengan model 25 titik basis keterlambatan keberangkatan paling tinggi diprediksi terjadi pada tanggal 4 April 2008 dan paling rendah terjadi pada 16 April 2008.
- Pada bulan April 2008 maskapai United Airlines akan mengalami delay keberangkatan paling rendah pada tanggal 17 dan delay keberangkatan paling tinggi akan terjadi pada tanggal 10. Jika dilihat berdasarkan hari, maskapai United Airlines diprediksi akan mengalami delay keberangkatan paling rendah pada hari Selasa, Kamis dan Jumat. Sedangkan delay keberangkatan paling tinggi terjadi pada hari Minggu.
- Pada bulan April 2008 hari Jumat dan Minggu diprediksi menjadi hari dengan delay keberangkatan paling tinggi dan hari Rabu merupakan hari dengan delay keberangkatan paling rendah.

REFERENCES

1. Micci-Barreca, D. (2001). A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems. SIGKDD Explor. Newsl., 3(1), 27–32.
2. Beckers, Thomas. (2021). An Introduction to Gaussian Process Models. arXiv preprint arXiv:2102.05497..

KELOMPOK 4: Matias J. (10119027) Susanty S. (10119060) Fadhilah S. (10119085) Yoshua R. (10119102) Andika Z. (10119111)